Research Methods in computer science
Spring 2024

Lecture 21

Omprakash Gnawali
April 3, 2024
Agenda

Conference updates
Paper feedback
Idea generation
Artifact comparison experiments
HW10
Paper review template

Summary
Strengths
Weaknesses
Detailed Comments
Paper feedback

Baseline for comparison
What questions need to be answered?
Clarify Claims/contributions
Generating Research Ideas

“Standing on the shoulders of giants”

Most ideas may not be new

New may be subjective

Adding a layer to an existing deep learning architecture

When is it new?

When is it not new?
Idea Generator Heuristics

Combination / Hybrid techniques
   From the same discipline
     (e.g., ....)
   From a different discipline
     (e.g., ....)

Address Gap/limitation (Incremental?)
   Handle some cases that were not handled
   Improve some (partial) aspects of dimension

Apply different datasets / settings / contexts
In-class group activity

Pick a paper

Generate at least two derivative ideas

Present: original and derivative ideas
CS Experiments Today

Artifact Comparison Experiments

- Run the new artifact
- Run best-known prior work
- Compare

Simulations + “Real” experiments
Wireless Experiments Today

Protocol Comparison Experiments
- Run the new protocol
- Run best-known prior work
- Compare

Simulations + Testbed experiments
Serial Experiments

Run one protocol at a time

Compare the results

Difficult to distinguish the contribution of these variables

Environment

Protocol mechanisms
Concurrent Experiments

Run multiple protocols concurrently
Compare the results

Advantages
Consistent environment for both the protocols

Concerns
Contention of different types
Results from Serial CTP vs LQI Experiment on Tutornet

<table>
<thead>
<tr>
<th>Metric</th>
<th>LQI</th>
<th>CTP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Delivery</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cost</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Path Length</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Churn/node-hr</td>
<td>5.5</td>
<td>0.5</td>
</tr>
</tbody>
</table>
### Results from Concurrent CTP vs LQI Experiment on Tutornet

<table>
<thead>
<tr>
<th>Metric</th>
<th>CTP</th>
<th>LQI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Delivery</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Cost</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Path Length</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Churn/node-hour</td>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

The diagram shows a comparison of Concurrent CTP and LQI in terms of delivery, cost, path length, and churn/node-hour. The CTP consistently outperforms LQI in all metrics except delivery.
Putting Concurrent Methodology to Use: Expts. with External Interference

Engineered Scenario

Both protocols *struggle* in the same environment.
Putting Concurrent Methodology to Use: Experiments in a Dynamic Network

CTP and LQI react differently to dynamics.
Uncontrolled environment does not imply we cannot do fair comparisons
Level of Details
At What Level of Detail?

Descriptions

System and algorithm
Experiments
Datasets
Results
We use all available nodes in every experiment. In some testbeds, this means the set of nodes across experiments is almost but not completely identical, due to backchannel connectivity issues. However, we do not prune problem nodes. In the case of Motelab, this approach greatly affects the computed average performance, as some nodes are barely connected to the rest of the network.
5.1 Methodology

We conducted our experiments on a tiered network testbed with several Stargate nodes and 40 TelosB motes. All nodes are located above the false ceiling across multiple rooms and hallways on a floor of a large office building. The wireless environment above the false ceiling is harsh, with some links experiencing above 30% packet loss rates. All nodes run the Tenet stack modified to support AEM. In most experiments, we use a single Tenet master node. We configured the mote radios to transmit at -8.906 dBm, which results in a tree with 4-hop depth.
Experimental Methodology and Metrics  We now compare the performance of Tenet-PEG and mote-PEG. Our experiments are conducted on the testbed shown in Figure 7. This testbed consists of 56 Tmotes and 6 Stargates deployed above the false ceiling of a single floor of a large office building. The Stargate and mote radios are assigned non-interfering channels. This testbed represents a realistic setting for examining network performance as well as for evaluating PEGs. The false ceiling is heavily obstructed, so the wireless communication that we see is representative of harsh environments. The environment is also visually obstructed, and thus resembles say, a building after a disaster, in which a pursuit-evasion sensor network might aid the robotic search for survivors.
Results from the same Testbed
Network Metric

Converting these subjective descriptions to a more quantitative description
END and CTP Performance

![Graphs showing END and CTP Performance]
“We evaluate the throughput and delay benefits of CQIC using the Google Nexus device to download content from a Google server via a popular cellular network provider. Reflecting a common CDN scenario, this server is located near the network of the mobile carrier such that the cellular channel is the bottleneck link...”

[Lu 2015]
AI/ML/NLP

Many times standardized datasets or tasks compare systems in the same dataset

Tradition of shared notebooks/repo online

Faithful implementation of prior work often less challenging in systems areas but not entirely if related to operations/systems aspect of AI
Typical Expt. in NLP-related areas

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Char-CNN</th>
<th></th>
<th></th>
<th>BERT</th>
<th></th>
<th></th>
<th></th>
<th>SOTA</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Acc(%)</td>
<td>F1 (%)</td>
<td>Acc(%)</td>
<td>F1 (%)</td>
<td>Acc(%)</td>
<td>F1 (%)</td>
<td>Acc(%)</td>
<td>F1 (%)</td>
<td>Acc(%)</td>
<td>F1 (%)</td>
</tr>
<tr>
<td>PHEME</td>
<td>80.72</td>
<td>81.43</td>
<td>83.82</td>
<td>78.51</td>
<td><strong>86.41</strong></td>
<td>81.72</td>
<td>85.21</td>
<td><strong>82.74</strong></td>
<td>–</td>
<td>77.40</td>
</tr>
<tr>
<td>Liar</td>
<td>64.80</td>
<td>54.40</td>
<td>68.75</td>
<td>62.57</td>
<td>67.01</td>
<td>59.34</td>
<td><strong>70.72</strong></td>
<td><strong>62.60</strong></td>
<td>65.54</td>
<td>60.80</td>
</tr>
<tr>
<td>FNN-Gossipcop</td>
<td>78.59</td>
<td>55.30</td>
<td>80.58</td>
<td>57.67</td>
<td><strong>86.11</strong></td>
<td>68.10</td>
<td>85.69</td>
<td>66.70</td>
<td>80.80</td>
<td><strong>75.50</strong></td>
</tr>
<tr>
<td>FNN-Politifact</td>
<td>71.70</td>
<td>58.33</td>
<td>73.58</td>
<td>68.69</td>
<td>81.46</td>
<td>77.43</td>
<td>81.76</td>
<td>77.91</td>
<td><strong>90.40</strong></td>
<td><strong>92.80</strong></td>
</tr>
<tr>
<td>Rashkin-Politifact</td>
<td>88.34</td>
<td>82.82</td>
<td><strong>95.23</strong></td>
<td><strong>93.46</strong></td>
<td>88.62</td>
<td>84.92</td>
<td>94.66</td>
<td>92.46</td>
<td>–</td>
<td>56.00</td>
</tr>
<tr>
<td>Rashkin-Newsfiles</td>
<td>97.81</td>
<td>98.25</td>
<td>96.42</td>
<td>97.15</td>
<td><strong>99.64</strong></td>
<td><strong>99.71</strong></td>
<td>99.43</td>
<td>99.56</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>COVID-Zenodo</td>
<td>96.04</td>
<td>92.55</td>
<td>95.78</td>
<td>97.77</td>
<td><strong>97.45</strong></td>
<td><strong>98.66</strong></td>
<td>97.21</td>
<td>98.53</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>COVID-AAAI</td>
<td>89.39</td>
<td>88.67</td>
<td>89.62</td>
<td>89.03</td>
<td><strong>95.42</strong></td>
<td>95.07</td>
<td>95.20</td>
<td>94.68</td>
<td>–</td>
<td><strong>98.37</strong></td>
</tr>
<tr>
<td>ENRON email spam</td>
<td>97.64</td>
<td>97.67</td>
<td>97.90</td>
<td>98.43</td>
<td>99.33</td>
<td>99.32</td>
<td><strong>99.43</strong></td>
<td><strong>99.46</strong></td>
<td>95.88</td>
<td>95.76</td>
</tr>
<tr>
<td>SMS Spam</td>
<td>92.82</td>
<td>77.78</td>
<td>97.12</td>
<td>91.40</td>
<td>98.32</td>
<td>93.56</td>
<td><strong>98.42</strong></td>
<td><strong>94.06</strong></td>
<td>97.64</td>
<td>–</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>89.98</strong></td>
<td><strong>89.53</strong></td>
<td><strong>90.42</strong></td>
<td><strong>90.27</strong></td>
<td><strong>92.72</strong></td>
<td><strong>92.50</strong></td>
<td>93.42</td>
<td>93.22</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>
Datasets not always standardized

Describe the data in enough detail even if the dataset cannot be released to the public
DeepFace: Closing the Gap to Human-Level Performance in Face Verification
[Taigman 2014]

The SFC dataset includes 4.4 million labeled faces from 4,030 people each with 800 to 1200 faces, where the most recent 5% of face images of each identity are left out for testing. This is done according to the images’ time-stamp in order to simulate continuous identification through aging. The large number of images per person provides a unique opportunity for learning the invariance needed for the core problem of face recognition...

“See the supplementary material for more details about SFC.”
Supplementary Material:
DeepFace: Closing the Gap to Human-Level Performance in Face Verification

Yaniv Taigman       Ming Yang       Marc’Aurelio Ranzato       Lior Wolf
Facebook AI Research
Menlo Park, CA, USA
{yaniv, mingyang, ranzato}@fb.com

Tel Aviv University
Tel Aviv, Israel
wolf@cs.tau.ac.il
HW10

Full paper submission

Due: April 7